COOLEY LLP
BOBBY GHAJAR (198719)
(bghajar@cooley.com)
COLETTE GHAZARIAN (322235)
(cghazarian@cooley.com)
1333 2nd Street, Suite 400
Santa Monica, California 90401
Telephone: (310) 883-6400

MARK WEINSTEIN (193043)
(mweinstein@cooley.com)
KATHLEEN HARTNETT (314267)
(khartnett@cooley.com)
JUDD LAUTER (290945)
(jlauter@cooley.com)
ELIZABETH L. STAMESHKIN (260865)
(lstameshkin@cooley.com)
3175 Hanover Street
Palo Alto, CA 94304-1130
Telephone: (650) 843-5000

CLEARY GOTTLIEB STEEN & HAMILTON LLP
ANGELA L. DUNNING (212047)
(adunning@cgsh.com)
1841 Page Mill Road, Suite 250
Palo Alto, CA 94304
Telephone: (650) 815-4131

PAUL, WEISS, RIFKIN, WHARTON & GARRISON LLP
KANNON K. SHANMUGAM (*pro hac vice*)
(kshanmugam@paulweiss.com)
2001 K Street, NW
Washington, DC 20006
Telephone: (202) 223-7300

Counsel for Defendant Meta Platforms, Inc.

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN FRANCISCO DIVISION

RICHARD KADREY, *et al.*, Individual and Representative Plaintiffs, v. META PLATFORMS, INC., a Delaware corporation; Defendant.	Case No. 3:23-cv-03417-VC-TSH **DECLARATION OF NIKOLAY BASHLYKOV IN SUPPORT OF META'S MOTION FOR PARTIAL SUMMARY JUDGMENT**

I, Nikolay Bashlykov, declare:

1. I am over the age of 18 and am competent to make this declaration. I am a Research Engineer in the Generative AI ("Gen AI") division of Meta Platforms, Inc. ("Meta"). I have been employed by Facebook UK Limited, a subsidiary of Meta, since August 2022. I have personal knowledge of the facts contained in this declaration in support of Defendant Meta Platform Inc.'s Motion for Partial Summary Judgment. I declare that the following is true to the best of my knowledge, information, and belief, and that if called upon to testify, I could and would testify to the following.

Professional Background

2. In 2012, I received a specialist degree, which is a five year degree, in Computational and Applied Mathematics at Lomonosov Moscow State University in Moscow, Russia. I also received a Master's degree in Business Administration and Management from the University of Mannheim in Mannheim, Germany in 2014.

3. For more than 10 years, I have worked as an engineer, including at Ernst & Young, Wheely Ltd., and Meta. I began working on software engineering relating to machine learning at Ernst & Young, where I was a Team Lead for Machine Learning beginning in 2018. At Wheely Ltd., which is a vehicle for hire company based in London, I led a team of data engineers and machine learning scientists. In August of 2022, I joined Meta as a Research Engineer.

Downloading of Datasets

4. In my professional role at Meta, I, along with my team, participated in obtaining and processing datasets for the Llama 3 model, including data from a source called "Library Genesis" or "Libgen" in or around Spring of 2023. In all instances, our team's work was conducted as part of our job responsibilities at Meta. The Libgen dataset comprised three portions: "Fiction," "Scitech," and "Scimag." Based on available records and documentation Fiction and Scitech each contain books, and Scimag contains academic articles and publications rather than books. The direct download method was used to obtain the Fiction and Scitech files, meaning that they were downloaded directly rather than via a torrent or other peer to peer client.

5. The only portion of Libgen our team acquired during this 2023 time frame via a

torrent client was Scimag, which, to our team's understanding, did not contain any books. An automated script was employed to remove the downloaded files shortly after completion—typically within 60 seconds—to minimize any potential for post-download seeding within the torrent network.

6. The datasets were evaluated for their overall size and diversity. The selection process did not involve reviewing individual book titles or articles, but rather focused on general content categories. It was my understanding that assessing specific titles was not a standard criterion when determining the suitability of training data for the Llama models. To the best of my knowledge, neither I nor anyone on our team was aware at the time of download that the files from the Fiction and Scitech portions of Libgen contained any works by the plaintiffs in this lawsuit.

Removal of Duplicative Text

7. In connection with processing the datasets for use in Llama 3, including the Libgen books data from Fiction and Scitech discussed above, our team undertook efforts to clean the data. As part of these cleaning efforts, the team developed a script to remove specified categories of text data from the Libgen books datasets. These categories included (1) potential personally identifiable information ("PII") such as email addresses, (2) excessive new line characters (i.e., empty rows), (3) rows containing the words or symbols "ISBN," "copyright," "©," "All rights reserved," and "DOI" in the first or last 25% of the file, and (4) excessively repetitive text (i.e., lines with a small number of unique words). These actions were carried out as part of the team's general practice for cleaning datasets—a practice that aligns with standard industry approaches for training large language models.

8. In my experience, it is a common practice to remove repetitive text from LLM training data because this improves the quality of the text for training purposes. In other words, a model trained on data containing a large amount of repetitive text would likely perform more poorly than the same model trained on the same data, but with repetitive text removed.

9. Data from the third category was excluded from further processing because it primarily consisted of duplicative, boilerplate text common to most published books, and, based on

our team's understanding, would not include tokens considered useful for training a large language model. In fact, inclusion of this type of highly repetitive text may cause the model to "over-fit" to it and could negatively impact model performance. The team developed this script to remove duplicative data with the objective of enhancing the overall quality of the datasets, and ultimately, improving the performance of the Llama models trained on them.

10. In developing and using this script to remove certain repetitive text (categories 3 and 4 listed above), blank space (category 2), and PII (category 1) data from the Libgen books data, the team had no intent to hide or obscure any information about the books in the dataset or other content of training data used by Meta, including that Meta was using portions of Libgen (or any books contained in Libgen), or the copyright status of any books used for training.

11. From a technical standpoint, I have no basis to conclude that the removal of text using this script could have the effect of concealing information about Meta's training data. In other words, there is no evidence to suggest that the removal of text in the categories described would obscure the fact that any particular book was used to train Llama 3 compared to the scenario in which such repetitive information was retained.

12. In discussions with colleagues at Meta regarding this script, I have not encountered any indication that it was used with the intent of concealing Llama's training data.

13. The data cleaning described above was undertaken in accordance with standard industry practices for training large language models, with the objective of enhancing the quality of the data used for Llama 3 and promoting improved model performance.

I declare under penalty of perjury that the foregoing is true and correct. Executed on this 19 day of March, at 18:00.

/s/
Nikolay Bashlykov